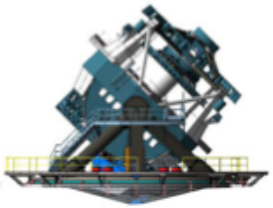
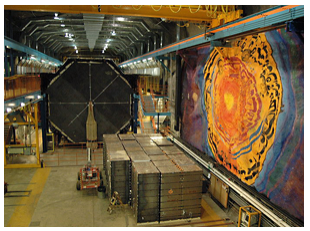
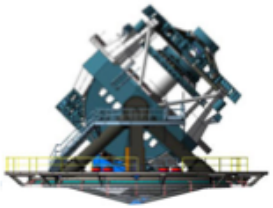


Cosmic Frontiers

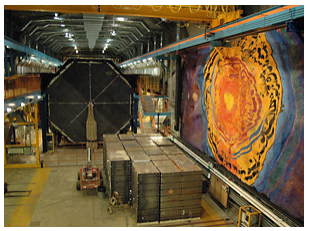


Alex Szalay, Andy Connolly, Salman Habib

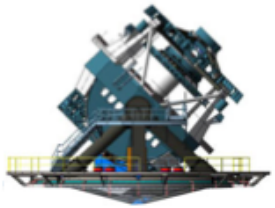
- Direct dark matter detection
- Novel Probes of Gravity and Dark Energy
- Dark Energy Facilities
- Distances
- Growth of Cosmic Structure: Probing Dark Energy Beyond Expansion
- Cross Correlations: Exploiting Multiple Probes, Surveys, and Techniques
- Spectroscopic Needs for Imaging Dark Energy Experiments



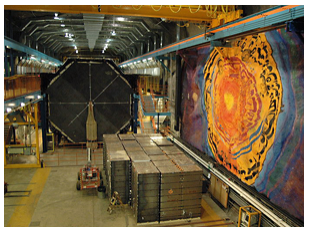
Cosmic Frontiers



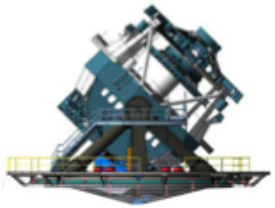
- Conclusions
  - There is a continued growth in data from Cosmic Frontiers experiments (currently exceed 1 PB, 50 PB in 10 years, 400 PB per year in 10-20 years).
  - Computational resources will have to grow to match the associated data rates (data intensive not just compute)
  - Data preservation and archiving (including the development of data storage and archiving technologies)
  - Infrastructure for data analytics applicable to large and small scale experiments will need to grow over the next decade (with an emphasis on sustainable software).



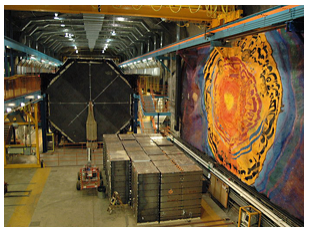
Cosmic Frontiers



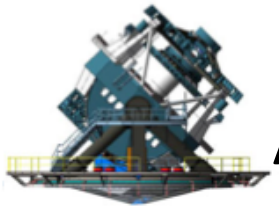
- Conclusions (cont'd)
  - Simulations (cosmological and instrument) will play a critical role in evaluating and interpreting the capabilities of current and planned experiments.
  - Require new computational models for distributed computing (including many-core systems)
  - Career paths (including tenure stream) for researchers who work at the forefront of computation techniques and science and the training of the next generation of researchers are critical to data intensive cosmology



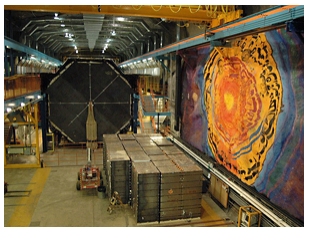
Cosmic Frontiers Meetings



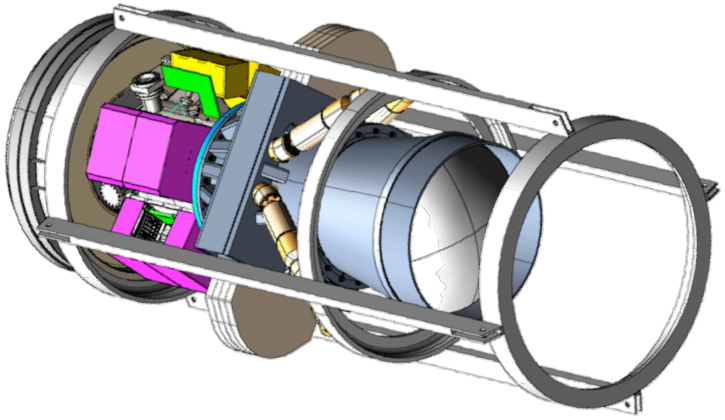
- Cosmic Frontiers Workshop (SLAC, 6-8 March)
  - direct and indirect dark matter detection
  - the cosmic microwave background
  - dark energy
  - complementarity of information obtained from colliders and astrophysical observations
  - cosmic particles and fundamental physics
- Contributors: Szalay, Myers, Morales, Borrill, Habib, Connolly, Nord, Gnedin, Schneider, O'Shea, Rhodes, Borgland, Jacobsen

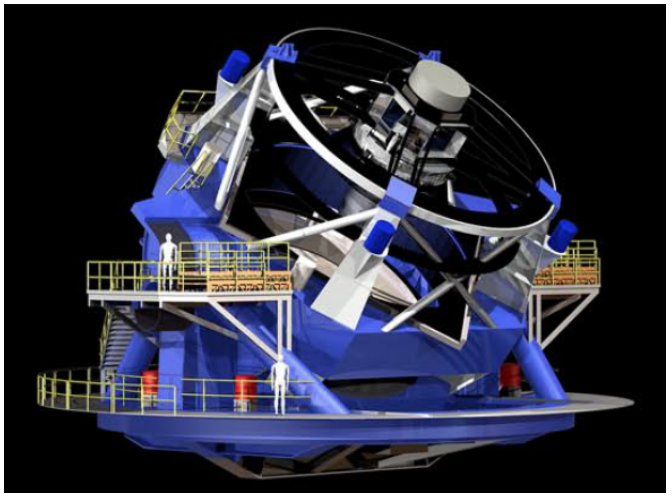


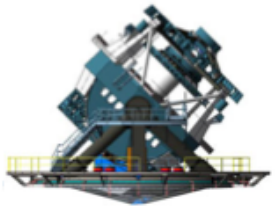
A decade of data: DES to LSST



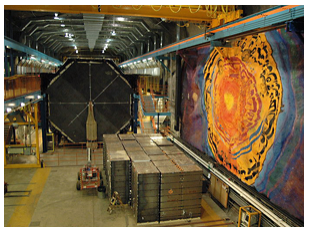
- Wide field and deep
  - DES: 5,000 sq degrees
  - LSST: 20,000 sq degrees
- Broad range of science
  - Dark energy, dark matter
  - Transient universe
- Timeline and data
  - 2012-16 (DES)
  - 2020 – 2030 (LSST)
  - 100TB - 1PB (DES)
  - 10PB - 100 PB (LSST)





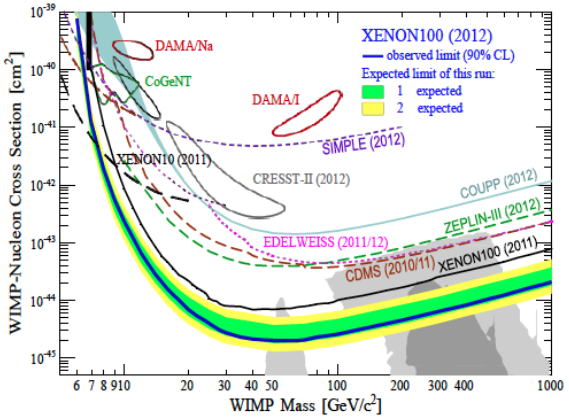


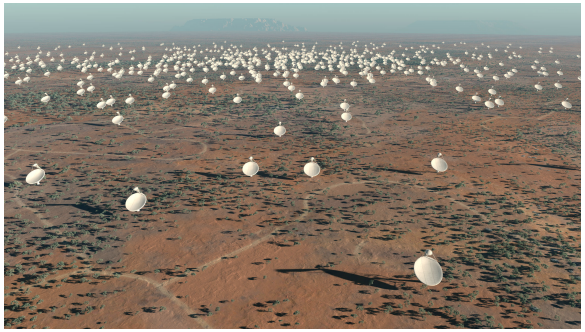
Growing volumes and complexity

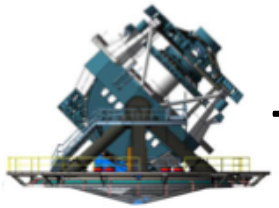


- CMB and radio cosmology
  - CMB-S4 experiment's 10^{15} samples (late-2020's)
  - Murchison Wide-Field array (2013-)
    - 15.8 GB/s processed to 400 MB/s
  - Square Kilometer Array (2020+)
    - PB/s to correlators to synthesize images
    - 300-1500 PB per year storage
- Direct dark matter detection
  - Order of magnitude larger detectors
  - G2 experiments will grow to PB in size

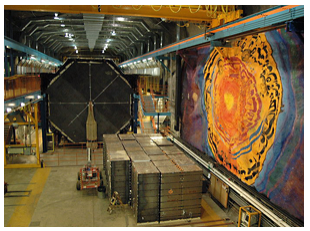
E. Aprile et al. (2012)



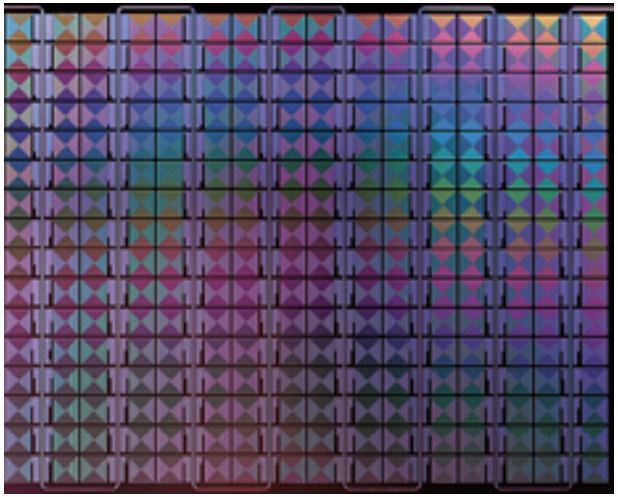


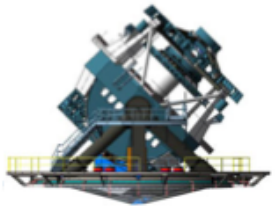


Technology developments

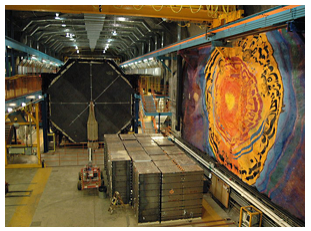


- Microwave Kinetic Inductance Detectors (MKIDs)
  - Energy resolving detectors (extended to optical and UV)
  - Resolving power: 30 < R < 150 (~5 nm resolution)
  - Coverage: 350nm – 1.3 microns
  - Count rate: few thousand counts/s
  - 32 spectral elements for uv/optical/ir photons

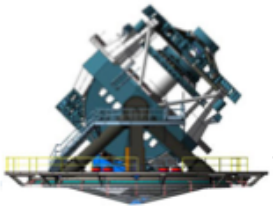




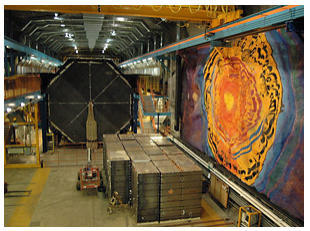
Data Management and Processing



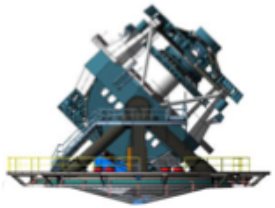
- Example use case from LSST (2020)
- The data volume associated with primary cadence
  - one 6-gigabyte image every 17 seconds
  - 15 terabytes of raw scientific image data / night
  - 100-petabyte final image data archive
  - 20-petabyte final database catalog
  - 2 million real time events per night, every night for 10 years (characterized within 60s)
- Reprocessing after 5 years (30PB)



Storage and archive technology

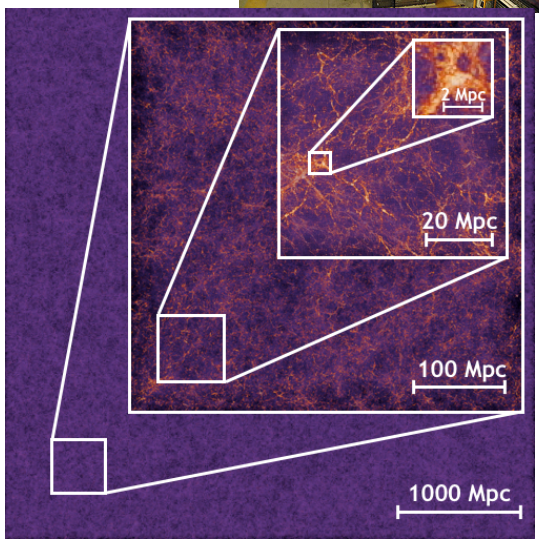


- Biggest archives to date 100-200 TB
- Large Synoptic Survey Telescope archive 5-60 PB
- Off the shelf solutions don't always match needs
  - SDSS used SQLServer but built HTM on top
  - Google, Facebook, Amazon build their own systems
- Usage patterns typically want access to calibrated data – occasional access to raw pixels
- Most operations are data intensive but today's supercomputers are not IO intensive
- IO major challenge (10 days: 1PB on 10 Gbps backbone)

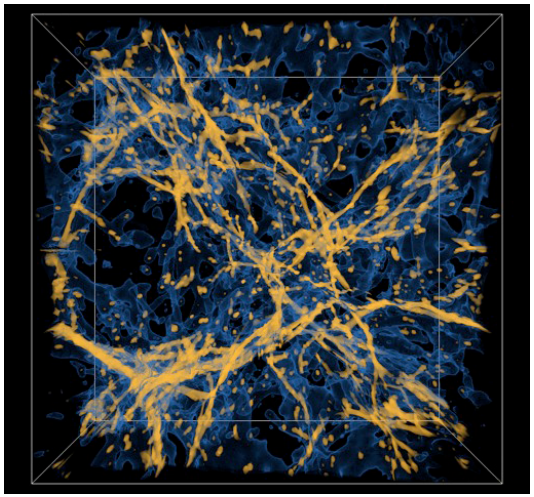


Role of Simulations in Cosmology

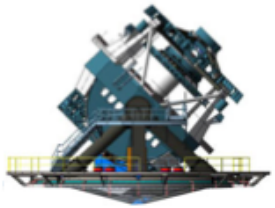
- Cosmological simulations are a basic tool for theory/modeling:
- Precision cosmology = statistical inverse problem
- Simulations provide predictions as well as means for estimating errors (MCMC campaigns)
- Synthetic catalogs play important roles in testing and optimizing data management and analysis pipelines and in mission optimization
- Roles in Future Missions:
- Theory: 1) BAO/P(k)/RSD/Shear/etc. predictions, 2) Error propagation through nonlinear processes (e.g., reconstruction, FoG compression), 3) Covariance matrices (many realizations, cosmological dependence?)
- Pipeline Validation
- Survey Design
- Simulation Type:
- N-body (gravity-only) simulations/derived products (e.g., simulated galaxy catalogs, emulators)
- Hydro simulations/derived products (e.g., Ly-alpha P(k))



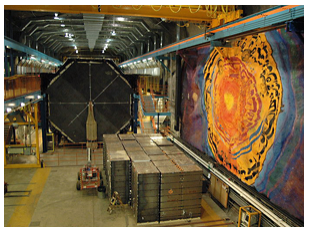
N-body: 1.07 trillion particle run with HACC



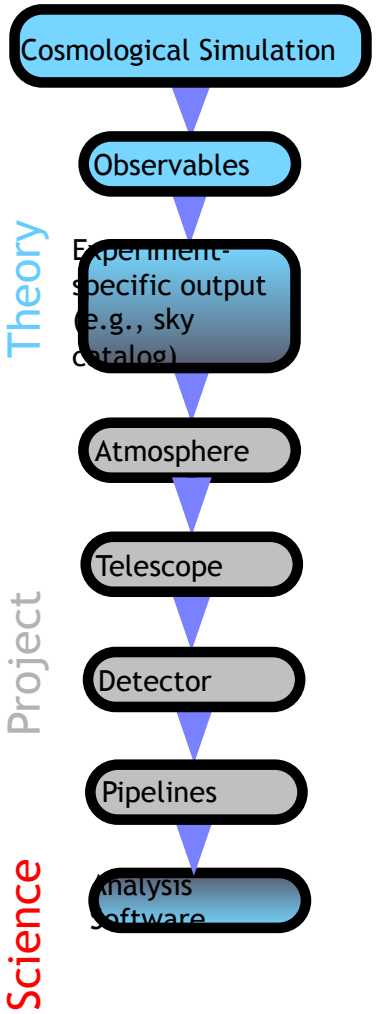
'Hydro' Simulations: Ly-alpha box with Nyx



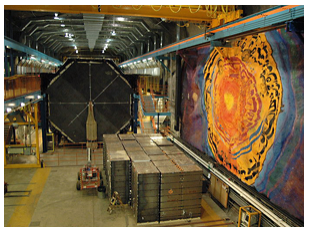
Cosmological Simulation Goals 1: Catalogs



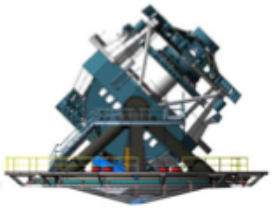
- Synthetic Catalogs:
- Generate a 'mock' skies that mimic particular surveys -- should be consistent with current observations and have sufficient realism, accuracy goals can be relatively relaxed
- Different catalogs for different purposes, evolving over time (depth/ breadth trade-offs, realism, schedule matching)
- Key Issues:
- How to model targets (different types of galaxies, QSOs, Ly-alpha forest, --)
- Role of simple(r) halo-based models vs. hydro simulations and SAMs, significant work needed to develop consistent approaches
- Number of cosmologies and realizations needed
- Ability to represent multiple probes for cross-correlation studies (CMB/galaxy distribution, Ly-alpha/QSOs, --)
- Training and validation -- close interaction with observers needed



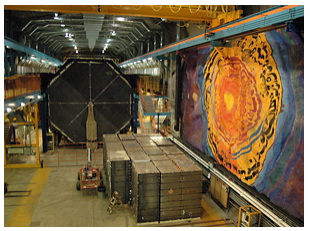
Cosmological Simulation Goals 2: Inference



- Theory/Modeling of Cosmological Probes:
- Model different probes at required resolution/area/accuracy/depth; develop ability for fast predictions at specified parameter settings for cosmological and physical/modeling parameters (emulators, simplified models)
- Computational infrastructure for inverse analysis (MCMC or other MC methods, data compression, emulators/simplified models, covariance matrices, --)
- Key Issues:
- Development of N-body codes -- ability to scale over survey sizes and resolve small-scale structure; enhance range of physics treated (modified gravity, model missing physics)
- Hydro codes and feedback with applications to galaxies/groups/clusters, Ly-alpha forest, reionization; studies of 'baryonic' systematic errors
- Post-processing of simulation data presents a major data-intensive computing challenge; management and serving of simulation data and databases



Computational Infrastructure

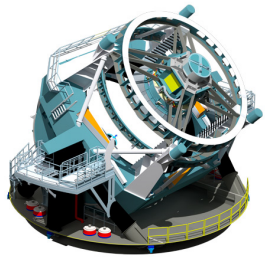


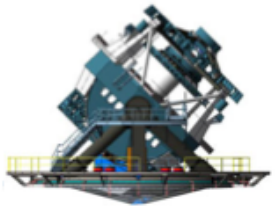
- High Performance Computing:
- Adapt to ongoing and future architectural changes (different types of complex nodes, end of naive weak scaling as memory/ core reduces, communication bottlenecks, multi-level memory hierarchy, power restrictions, --)
- Advent of new programming models -- how to rewrite large code bases?
- Data Archiving and Serving:
- As simulation data products become larger and richer and the process for generating them more complex, data archives, databases, and facilities for post-analysis are becoming a pressing concern
- Development of powerful, easy-to-use remote analysis tools motivated by network bandwidth restictions on large-scale data motion (compute and data-intensive platforms will likely be co-located)

Supercomputers for simulation campaigns

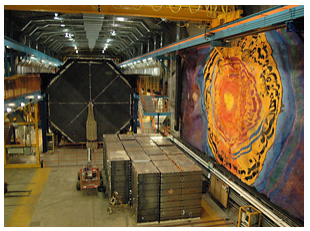
Data-Intensive Scalable Computing Systems for interactive analysis

Data Sources

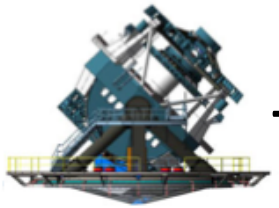




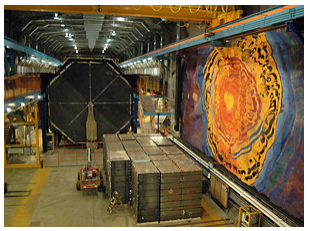
Cosmic Frontier Computing Requirements



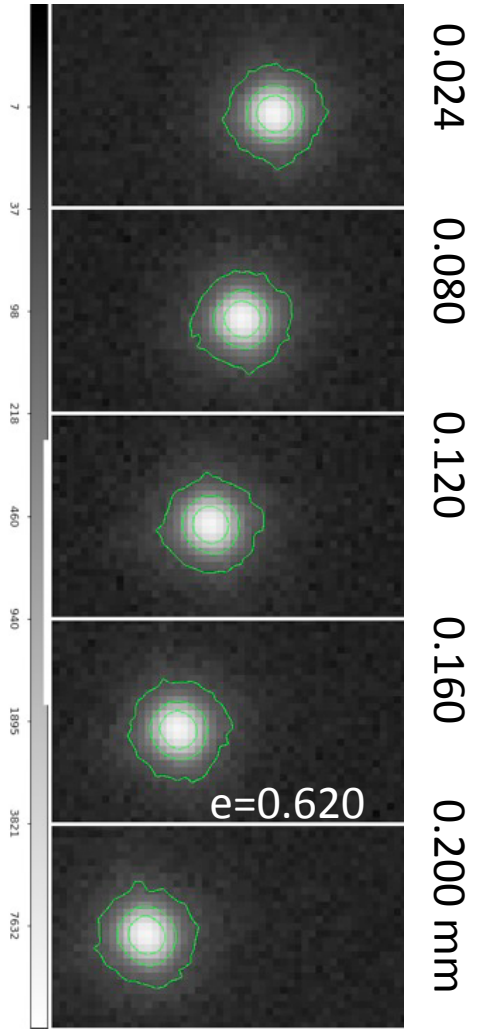
- Probes Based on Structure Formation:
- Most computationally intensive and connected to the detailed implementation of large-scale surveys (cutting-edge supercomputing, exascale targets)
- Associated with the largest and most complex datasets (many PBs)
- Indirect Dark Matter Searches:
- Simulation activity will partially mirror that of structure formation probes except at much higher spatial resolution
- Less intense data and observational modeling requirements (astrophysical backgrounds need to be understood, however)
- Other Significant Areas:
- Supernova simulations for understanding Sn Ia systematics/observations (computationally intensive, needs not fully established)
- Early Universe simulations for possible CMB/LSS signatures (computational needs typically modest)

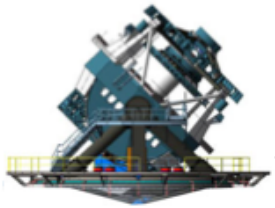


The need for instrument simulation

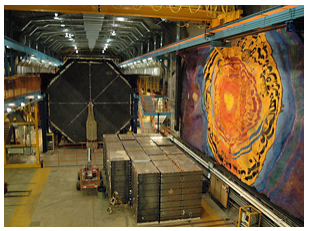


- Great successes in simulation
  - Fluka, Geant4, MCNPX, MUSUN, MUSIC and SOURCES.
  - Future experiments driven by systematics and how we can model and correct for them.
  - Instrument simulations provide these capabilities (recognition of this depends on the field)
  - Opportunity for shared resources and expertise



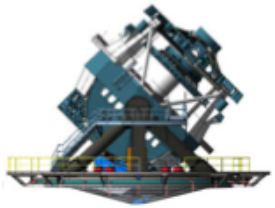


Scalable Analytics

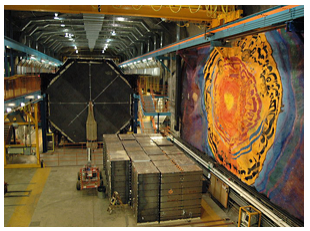


- Cosmologists willing to embrace new technologies/ techniques (SDSS introduced SQL – now ubiquitous)
- Success because researchers saw impact early
- No current way to do “science” within the database
- Need to move the analysis into the archive
- Needs sustainable (and reproducible software initiatives) software development not just “build your own” approaches

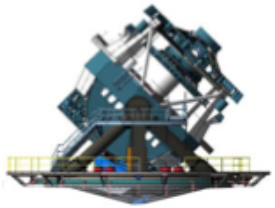
… what is the right mix of statistics/CS/physics to enable new science



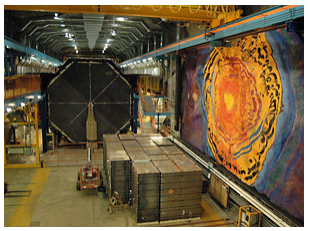
Open questions…..



- Many challenges are sociological – how do we change the way a field approaches data analysis
- Archives have been a successful model but we still "just" download data
- How do we (or should we) educate physicists to think like computer scientists?
- Parallelism (including multicore and Hadoop) can change the playing field but at the cost of significant development time
- Who do we support: public, typical users, power users



Conclusions



- Continued growth in data from Cosmic Frontiers (currently 1 PB, 50 PB in 10 yrs, 400 PB/yr in 20 years).
- Computational resources will have to grow to match the data rates (data intensive not just compute)
- Infrastructure for data analytics and sustainable software needs will grow over the next decade
- Simulations (cosmological and instrument) will play a critical role in evaluating and interpreting experiments
- Data preservation and archiving needs to scale
- New computational models for distributed computing
- Training and career paths (including tenure stream) for researchers who work at the forefront of computation techniques and science is critical